doi:10.1093/nar/gks489

CombFunc: predicting protein function using heterogeneous data sources

Mark N. Wass[1,2,]*, Geraint Barton[1] and Michael J. E. Sternberg[1,]*

[1]Centre for Bioinformatics, Imperial College London, London, SW7 2AZ, UK and [2]Structural Computational Biology group, CNIO, Madrid, Spain

Received February 27, 2012; Revised April 29, 2012; Accepted May 8, 2012

ABSTRACT

Only a small fraction of known proteins have been functionally characterized, making protein function prediction essential to propose annotations for uncharacterized proteins. In recent years many function prediction methods have been developed using various sources of biological data from protein sequence and structure to gene expression data. Here we present the CombFunc web server, which makes Gene Ontology (GO)-based protein function predictions. CombFunc incorporates ConFunc, our existing function prediction method, with other approaches for function prediction that use protein sequence, gene expression and protein–protein interaction data. In benchmarking on a set of 1686 proteins CombFunc obtains precision and recall of 0.71 and 0.64 respectively for gene ontology molecular function terms. For biological process GO terms precision of 0.74 and recall of 0.41 is obtained. CombFunc is available at http://www.sbg.bio.ic.ac.uk/combfunc.

INTRODUCTION

Protein function prediction is essential to provide insight to the functions of uncharacterized proteins. This is highlighted by the gap between the large number of proteins that have been identified and the small percentage of them that have been functionally characterized (1). Annotation transfer using BLAST (2) represents a standard and widely-used method of function prediction but as protein function is often only conserved by homologues sharing a high sequence identity, this approach can be prone to errors (3). In recent years many methods have been developed to improve upon BLAST-based annotation transfer. This has included methods such as GOtcha (4) and PFP/ESG (5,6), which combine the Gene Ontology (GO) (7) annotations present in multiple homologues and use their *e*-values to weight predictions or use machine learning to optimize predictions (8). Phylogenomics approaches distinguish between orthologues and paralogues to infer function (9). The presence of domains from Interpro (10) or Pfam (11) are used for electronic annotation in GO annotations (12) and combinations of domains have also been used for function prediction (13). In ConFunc we used conserved residues representative of individual GO terms to predict protein function (14). Other methods have used protein–protein interaction networks (15,16), gene co-expression (17,18) or multiple protein features including protein disorder and secondary structure (19).

Some methods combine predictions from multiple sources of data (20–26). This includes methods that use Bayesian approaches (24) or Support Vector Machines (SVMs) (23,25) to combine predictions. Some of these methods are available as web servers. The ProKnow (21) webserver combines the evidence from multiple sources to make overall predictions of GO functions. In contrast the ProFunc (20) and PredUS (22) servers do not make overall predictions of protein function, instead they enable the user to explore the results of the many sequence and structural analyses that they perform. Further details of these methods and others are available in recent reviews (1,27).

Here we present CombFunc a server for GO-based protein function prediction. CombFunc incorporates ConFunc (14) our existing sequence based function prediction method and it also extends our recent use of multiple methods to predict the functions of proteins in the *Plasmodium berghei* male gamete (28). CombFunc uses sequence information including BLAST/PSI-BLAST (29) annotation transfer, domain information from Interpro, protein–protein interaction data from IntAct (30) and MiNT (31) and gene expression data from COXPRESdb (32).

MATERIALS AND METHODS

The CombFunc algorithm

CombFunc obtains information from multiple analyses which are then combined using a SVM (33) to make an

*To whom correspondence should be addressed. Tel: +34 917 328 000; Email: mark.wass04@imperial.ac.uk
Correspondence may also be addressed to Michael J. E. Sternberg. Tel: +44 20 7594 5212; Fax: +44 20 7594 5264; Email: m.sternberg@imperial.ac.uk

overall prediction. The data sources used are described below.

The sequence-based sources of input to CombFunc are: ConFunc, BLAST/PSI-BLAST annotation transfer, domain information and a sequence search against the fold library of Phyre2 (34), our in-house protein structure prediction server. ConFunc is run as previously described in Wass and Sternberg (14). Both BLAST and PSI-BLAST are used to search for GO annotated homologues of the query sequence in UniProt (35). Where PSI-BLAST is used, UniRef50 is initially searched and the profile generated is used to search the full UniProt database as this approach has been shown to improve the identification of homologues (36). Domain information is obtained using Interpro (10) and Pfam domain combinations are also used to make predictions as described in (13). HHsearch (37) is used to search the fold library of Phyre2 to identify structures homologous to the query sequence, whose annotations are input to the SVM. All methods use only experimentally determined GO annotations.

The non-sequence-based data sources are protein–protein interactions (PPI) and gene co-expression. PPI data are obtained from both IntAct (30) and MiNT (31). Function prediction is performed by simple neighbour counting (38) and indirect neighbours are also included (15). Gene expression data is obtained from the COXPRESdb database (32), which contains expression data for Human, Mouse, Rat, Chicken, Zebrafish, Fly and Nematode. COXPRESdb uses a mutual rank score to determine the strength of co-expression, which is calculated as the geometric mean of the correlation rank of gene A to gene B and of gene B to gene A. The frequency of GO terms within the set of co-expressed genes with a mutual rank less than 50 (39) is input into the SVM.

CombFunc uses each of the individual methods to identify GO terms that may be associated with the query. Features associated with the GO terms identified by individual methods are used by CombFunc to make a final prediction of the query function. The features used for each method are listed in Supplementary Table S1 and are described below.

For BLAST and PSI-BLAST the top annotated hit is identified and the GO terms it is annotated with are used for prediction. The features from BLAST and PSI-BLAST include the *e*-value of the top annotated hit, the sequence identity between the query and top annotated hit and also the sequence coverage of the query by the top hit. Additionally for PSI-BLAST data the annotations of multiple sequences are considered by calculating the i-score as used in GOtcha (4). For terms identified by the interactome analysis the features correspond to the fraction of direct and also indirect neighbours that are annotated with that term. For terms present in the Interpro analysis, the feature corresponds to the lowest *e*-value of a domain hit annotated with that term (maximum of 1). For the Pfam domain combinations analysis the feature is 1 if predicted by the method and 0 otherwise. Features from the Phyre2 fold library use terms present in the top annotated hit and use the probability score from HHsearch (37) between the query and the hit and also the sequence coverage of the query by the hit. Features for GO terms identified from expression data use a number of features including: the fraction of co-expressed proteins annotated with the function and the minimum, average and maximum mutual rank and correlation coefficients of the co-expressed proteins. Finally a feature is included for each of the individual level 1 GO terms (i.e. binding and catalytic function in molecular function). These features are set to 1 if they are a parent term of the term being considered and zero otherwise.

CombFunc uses three classifiers for the molecular function and biological process categories. As the features associated with GO terms are likely to vary depending on their location in the GO graph, the three classifiers are used for different levels of GO. One classifier considers only terms one level below the root (e.g. catalytic activity or binding in the molecular function category), the second considers terms in the next two levels, while the third classifier considers all more specific terms. The scores output from the SVMs are converted to probabilities as described in Platt (40). The classification process is repeated 10 times, using the 10 sets of optimized SVMs generated during cross-validation. GO terms are predicted to be a function of the query protein if they are predicted to be so by at least 5 of the 10 sets of SVMs with a probability score set as an average of the probability scores for the SVMs that predicted the function.

Generating a test set

A test set of proteins with experimental GO annotations in both the molecular function and biological process GO categories was extracted using the UniProt-GOA annotations from December 2011. This was reduced to a representative set with less than or equal to 25% sequence identity using CD-HIT (41). Of the resulting 6686 sequences, 5000 were used for cross-validation and the remaining 1686 for final testing of the server.

SVM training

The SVMs were generated using SVMlight (33). A linear kernel was used for classification. For each of the 10-fold, eight were used for training, a further fold was used for optimization and the SVM tested on the remaining fold. In cross-validation each SVM was optimized for the trade off between training error and margin. As the training data is unbalanced with many more negative examples than positive ones we also assessed the effect of the cost factor to identify how training errors on positive examples should outweigh those on negative examples (see Supplementary Material section).

EVALUATING COMBFUNC PERFORMANCE

Here we assess the performance of CombFunc using the set of sequences that were not used in cross-validation. The performance of CombFunc on this set of 1686 sequences was assessed using precision and recall calculated as described in Wass and Sternberg (14). The precision-recall graphs in Figure 1 show the performance of CombFunc at a range of thresholds and a comparison

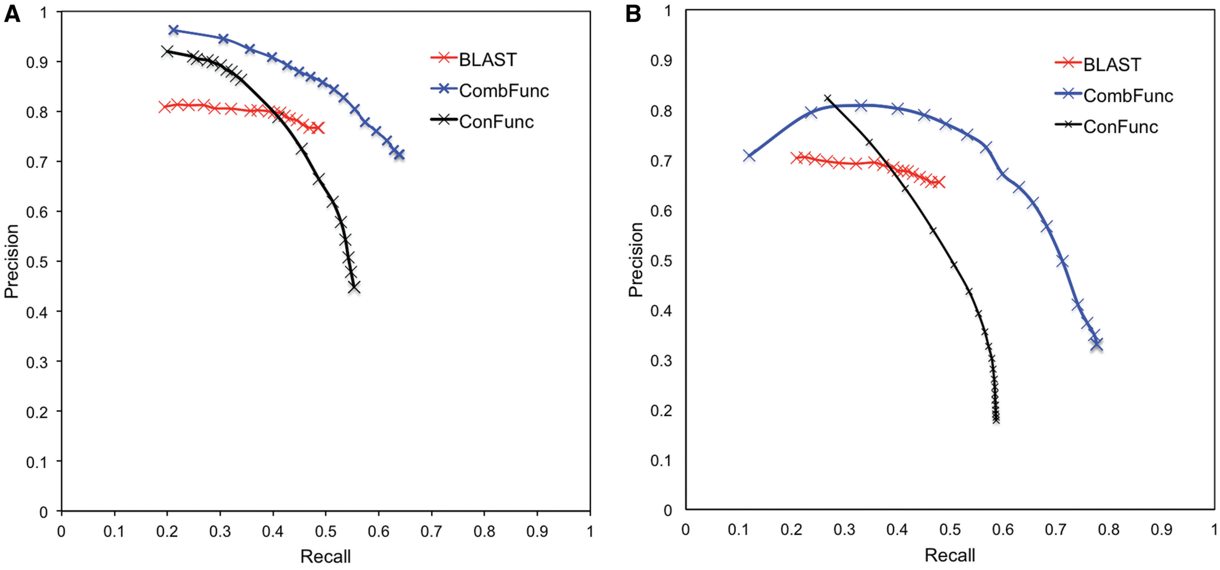


Figure 1. Benchmarking CombFunc. Precision-recall graphs showing the performance of CombFunc on 1686 sequences not used in cross-validation. CombFunc results are shown in blue, ConFunc in black and BLAST in red. For (**A**) the GO molecular function and (**B**) biological process categories.

with the performance of BLAST annotation transfer. For CombFunc the performance is assessed at confidence thresholds in the range 0–1. We observe that at high confidence (>0.95) CombFunc obtains high precision (0.96) and low recall (0.21). As the threshold is reduced the recall increases while precision reduces and including low confidence predictions CombFunc obtains precision and recall of 0.71 and 0.64 respectively (Figure 1A). CombFunc does not perform as well on biological process terms with both lower recall and precision at equivalent confidence scores. Using a confidence threshold of 0.3 obtains precision of 0.74 and recall of 0.41.

For comparison the performance of BLAST and ConFunc on the same dataset was considered. For BLAST (Version 2.219) annotation transfer the UniProt database (version December 2011) was searched and the annotation of the top (lowest *e*-value) experimentally annotated hit transferred to the query sequence. A range of precision and recall scores is obtained by only transferring the annotation if the top hit has an *e*-value below a threshold, which was varied from $0 - 1e^{-03}$. For ConFunc precision-recall values were obtained using a threshold for the ratio score (range 0–1). For benchmarking of all three methods, sequences with >99% sequence identity were excluded for the sequence based prediction components to ensure that the query sequence was not used to make predictions for itself.

We observe that CombFunc performs better than both BLAST and ConFunc. For ConFunc predictions there is a large reduction in precision as the prediction threshold is reduced. ConFunc considers all of the annotations that are present in the homologues of the query identified by BLAST. This often includes the annotation of the query sequence but additionally includes many other functions that are not annotations of the query sequence. At low thresholds many false positive predictions are made. In contrast through the use of multiple data sources and machine learning, CombFunc does not have such a large reduction in precision at lower thresholds, particularly when predicting molecular function terms (Figure 1).

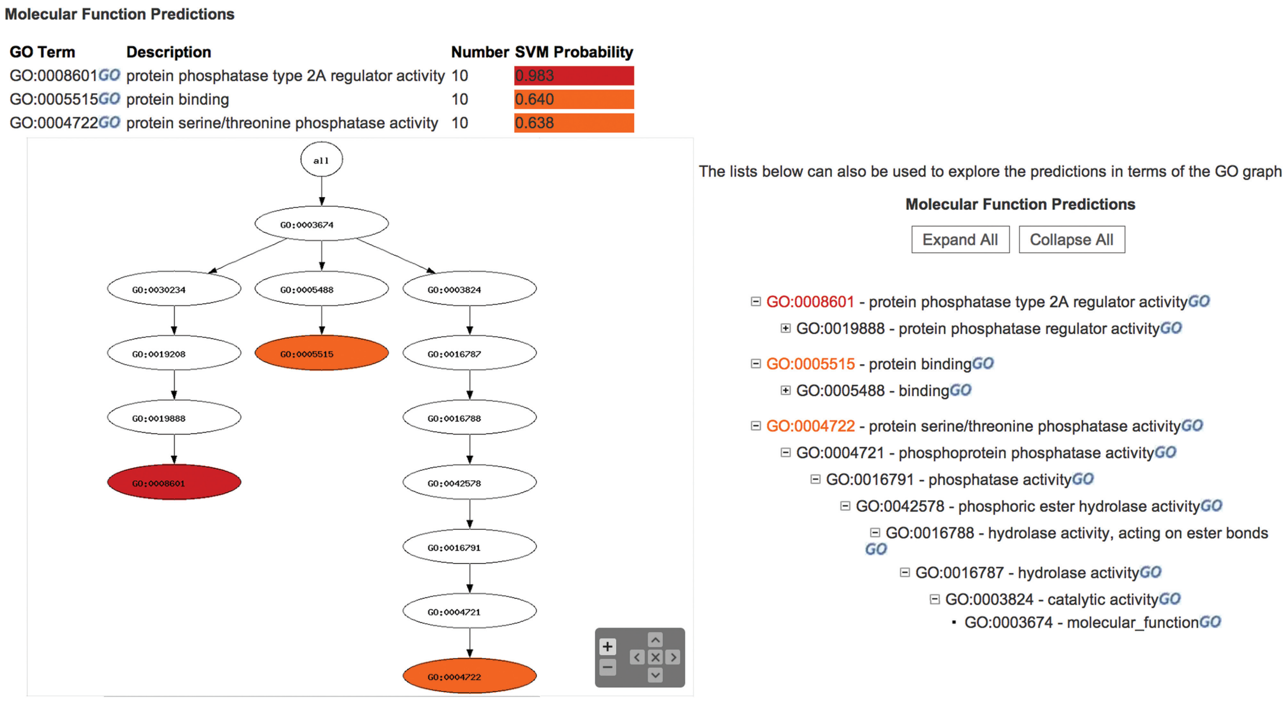
THE COMBFUNC WEB SERVER

CombFunc is available at http://www.sbg.bio.ic.ac.uk/combfunc. Users are required to submit a protein sequence in fasta format and they may also input the UniProt accession of the query sequence. The UniProt accession is required to perform the PPI and co-expression analyses. Processing time for each submission can vary from between 20 min to a few hours, this is largely due to the time taken to perform the search of the Phyre2 fold library.

Results output

CombFunc results output is split into two main sections. The prediction section provides details of the functions predicted by the SVM. In the second section details of the data generated from each of the individual analyses are provided, which users can explore to obtain further details of the data used to make the prediction.

The prediction section displays separate results for molecular function and biological process predictions. For both of these GO categories a table of the predictions lists the term, its name and the probability score of the prediction, this has a range of 0–1, with 1 being the highest confidence (Figure 2). The probability scores are colour coded to indicate the confidence of the predictions, ranging from yellow for low probability predictions to red for high probability. Longer descriptions of the predicted functions are displayed adjacent to the table when the mouse is moved over the rows of the table. Additionally links to the GO terms on the GO website are provided, enabling the user to access external further information about the predicted GO terms.



Molecular Function Predictions

GO Term	Description	Number	SVM Probability
GO:0008601GO	protein phosphatase type 2A regulator activity	10	0.983
GO:0005515GO	protein binding	10	0.640
GO:0004722GO	protein serine/threonine phosphatase activity	10	0.638

The lists below can also be used to explore the predictions in terms of the GO graph

Molecular Function Predictions

- ⊟ GO:0008601 - protein phosphatase type 2A regulator activityGO
  - ⊞ GO:0019888 - protein phosphatase regulator activityGO
- ⊟ GO:0005515 - protein bindingGO
  - ⊞ GO:0005488 - bindingGO
- ⊟ GO:0004722 - protein serine/threonine phosphatase activityGO
  - ⊟ GO:0004721 - phosphoprotein phosphatase activityGO
    - ⊟ GO:0016791 - phosphatase activityGO
      - ⊟ GO:0042578 - phosphoric ester hydrolase activityGO
        - ⊟ GO:0016788 - hydrolase activity, acting on ester bonds GO
          - ⊟ GO:0016787 - hydrolase activityGO
            - ⊟ GO:0003824 - catalytic activityGO
              - GO:0003674 - molecular_functionGO

Figure 2. Display of a CombFunc prediction. CombFunc predictions are displayed in a table showing the confidence of the prediction and in an image and list placing them in the context of GO structure.

The predictions are visualized within the GO graph in an image that displays a subgraph of GO containing all of the predicted terms and their parent terms (Figure 2). Again predicted terms are colour coded to indicate the confidence of their prediction. The image has a zoom function that enables users to zoom into different areas of the graph to investigate the predictions, which is particularly useful when multiple terms are predicted and the subgraph becomes large. Additionally, the predictions are displayed as an expandable list, which enables similar investigation of the predicted terms.

The second section of the results page contains the output from each of the individual analyses performed. The data associated with each analysis are initially hidden so that the user can view only the analyses they wish to. For each analysis a table lists the GO terms identified by the method and the values or scores associated with those terms. Interpro results are additionally displayed graphically enabling the user to identify the location of the hits on the query sequence. For all analyses the same colour coding as for the main predictions is used to give an indication of how 'good' the different scores displayed are. This includes colour coding sequence identity and *e*-values of BLAST hits and mutual rank values for gene co-expression. Where relevant, links to external data on the GO, UniProt and Intpero websites are provided.

For each submission to CombFunc a submission is also made to 3DLigandSite (42,43), our in-house ligand binding site prediction server. This enables users to combine the function prediction results with the binding site prediction of 3DLigandSite. A link to the 3DligandSite results is provided at the end of the analysis section.

CONCLUSION

CombFunc was developed to utilize the multiple data sources that are available for protein function prediction. In benchmarking CombFunc obtains good performance with 0.71 and 0.64 precision and recall respectively for molecular function GO terms and precision of 0.74 and recall of 0.41 for biological process terms. The CombFunc server provides a resource for users to view predicted functions in both tabular and graphical formats, access to the raw data from each individual method and access to external resources to enable users to explore the functions and data used to make predictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors would like to thank Lawrence Kelley for advice on the use of SVMs, Suhail Islam for technical support and Michael Tress for helpful discussions about function prediction.

FUNDING

Biotechnology and Biological Sciences Research Council [BB/F020481/1 to M.N.W.]. Funding for open access charge: Imperial College London Library.

Conflict of interest statement. M.J.E.S. is a founder director of Equinox Pharma Ltd, holds shares in the company, and has obtained remuneration from the company. Equinox Pharma Ltd is exploiting computational methods for drug discovery and markets software.

REFERENCES

1. Erdin,S., Lisewski,A.M. and Lichtarge,O. (2011) Protein function prediction: towards integration of similarity metrics. *Curr. Opin. Struct. Biol.*, **21**, 180–188.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Devos,D. and Valencia,A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.
4. Martin,D.M., Berriman,M. and Barton,G.J. (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC bioinformatics*, **5**, 178.
5. Hawkins,T., Chitale,M., Luban,S. and Kihara,D. (2009) PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins*, **74**, 566–582.
6. Chitale,M., Hawkins,T., Park,C. and Kihara,D. (2009) ESG: extended similarity group method for automated protein function prediction. *Bioinformatics*, **25**, 1739–1745.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
8. Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. *Proteins*, **79**, 2086–2096.
9. Engelhardt,B.E., Jordan,M.I., Srouji,J.R. and Brenner,S.E. (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Res.*, **21**, 1969–1980.
10. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
11. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
12. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
13. Forslund,K. and Sonnhammer,E.L. (2008) Predicting protein function from doma in content. *Bioinformatics*, **24**, 1681–1687.
14. Wass,M.N. and Sternberg,M.J. (2008) ConFunc–functional annotation in the twilight zone. *Bioinformatics*, **24**, 798–806.
15. Chua,H.N., Sung,W.K. and Wong,L. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**, 1623–1630.
16. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
17. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
18. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
19. Lobley,A., Swindells,M.B., Orengo,C.A. and Jones,D.T. (2007) Inferring function using patterns of native disorder in proteins. *PLoS Computat. Biol.*, **3**, e162.
20. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
21. Pal,D. and Eisenberg,D. (2005) Inference of protein function from protein structure. *Structure London*, **13**, 121–130.
22. Zhang,Q.C., Deng,L., Fisher,M., Guan,J., Honig,B. and Petrey,D. (2011) PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res.*, **39**, W283–W287.
23. Guan,Y., Myers,C.L., Hess,D.C., Barutcuoglu,Z., Caudy,A.A. and Troyanskaya,O.G. (2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol.*, **9(Suppl. 1)**, S3.
24. Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
25. Obozinski,G., Lanckriet,G., Grant,C., Jordan,M.I. and Noble,W.S. (2008) Consistent probabilistic outputs for protein function prediction. *Genome Biol.*, **9(Suppl. 1)**, S6.
26. Nariai,N., Kolaczyk,E.D. and Kasif,S. (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PloS One*, **2**, e337.
27. Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. *Brief. Funct. Genom Proteomics*, **7**, 291–302.
28. Wass,M.N., Stanway,R., Blagborough,A.M., Lal,K., Prieto,J.H., Raine,D., Sternberg,M.J., Talman,A.M., Tomley,F., Yates,J. *et al.* (2012) Proteomic analysis of Plasmodium in the mosquito: progress and pitfalls. *Parasitology*, February 16 (doi:10.1017/S0031182012000133; epub ahead of print).
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
31. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
32. Obayashi,T. and Kinoshita,K. (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res.*, **39**, D1016–D1022.
33. Vapnik,V.N. (1999) An overview of statistical learning theory. *IEEE Transact Neural Networks / a publication of the IEEE Neural Networks Council*, **10**, 988–999.
34. Kelley,L.A. and Sternberg,M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protocols*, **4**, 363–371.
35. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, **39**, D214–D219.
36. Chubb,D., Jefferys,B.R., Sternberg,M.J. and Kelley,L.A. (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics*, **26**, 2664–2671.
37. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
38. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.
39. Obayashi,T., Hayashi,S., Shibaoka,M., Saeki,M., Ohta,H. and Kinoshita,K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, **36**, D77–82.
40. Platt,J.C. (1999) *Advances in Large Margin Classifiers*, Vol. 1. MIT Press, Cambridge, Massachusetts, US, pp. 61–74.
41. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
42. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
43. Wass,M.N. and Sternberg,M.J. (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, **77(Suppl. 9)**, 147–151.